“A Review of Big Data Analytics in Healthcare”

Gautham Sriman Narayan

CSCI Y790 Independent Study

Fall 2015

(under Prof. Geoffrey Fox)

CONTENTS

Introduction:

Today, the Healthcare Industry is transitioning from paper to electronic records, and generates a huge amount of data from numerous sources in the form of medical databases, patient records, etc. Valuable and potentially life-saving patterns can be inferred from these data. Yet, “Big Data Analytics” is still largely constrained and unstructured in the domain of Healthcare.

Big Data Analytics refers to the process of selecting, exploring and modelling large amounts of data. This process has become an increasingly pervasive activity in all areas of medical science research. It has resulted in the discovery of useful hidden patterns from massive databases. Data Analytics problems are often solved using different approaches from both Computer Science, such as multidimensional databases, machine learning, soft computing and data visualization; and Statistics, including hypothesis testing, clustering, classification, and regression techniques. Analytics of Big Data helps institutions make critical decisions faster and with a greater degree of confidence, and lowers the uncertainty in decision process. The integration of Big Data Analytics into Healthcare can lead to the improved performance of Medical Decision Support Systems and can enable the tackling of new types of problems that have not been addressed before.

An immense challenge facing the healthcare industry today, is the provision of quality services – i.e. correct, timely diagnosis and effective treatment, at affordable costs. Also, the amount of data to be analyzed for diagnostic purposes is huge and at times unmanageable (“Big Data”). In this context, Big Data Analytics can be used to efficiently infer patterns and rules from earlier treatments, thus helping to make diagnosis more objective and reliable. There is a huge amount of untapped data which can be analyzed to obtain useful information through Analytics of Big Data and Data Mining.

According to a study, “In 2012, worldwide digital healthcare data was estimated to be equal to 500 petabytes and is expected to reach 25,000 petabytes in 2020”. Also, it’s estimated that in 2015, an average hospital is managing 665 terabytes of patient data, 80% of which is unstructured medical imaging data [5]. This goes to show the vast amount of medical data that is at the disposal to analyze and gather meaningful relationships for the betterment of the Healthcare Industry.

Here, in this report, I have surveyed Big Data Analytics in the sensitive and critical field of Healthcare.

In the next topic, I will cover the Big Data Services in Healthcare.

Big Data Services for Healthcare

Big Data can be useful in a broad range of ways in the Healthcare sector [8].

(1) Clinical Perspective:

(a) Details of Patients- like previous medical diagnosis information, query by MRN, etc.
(b) Episodes of care across patient visits.

(2) Operations Perspective:

(a) Optimization activities like Process turn-around time can be got (time spent waiting in healthcare facility, etc.)
(b) Throughput measurements can be obtained.

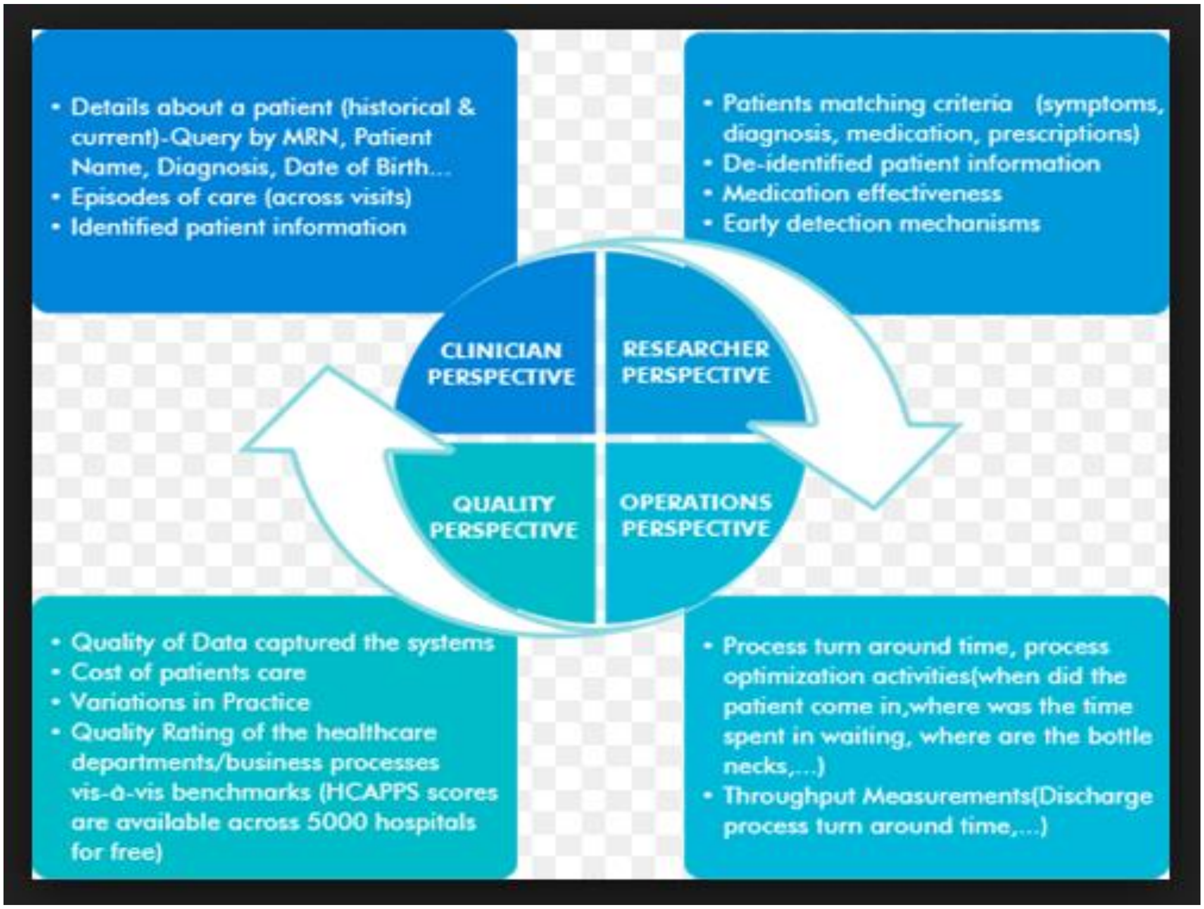
(3) Researcher Perspective:

(a) Medication efficiency and Early detection mechanisms.
(b) Diagnosis of Diseases based on symptoms, prescriptions, etc.

(4) Quality Perspective:

(a) Patient care cost.
(b) Effective alteration in medical practice,
(c) Test the quality of medical data.

In the subsequent section, we will see the Impact of Big Data on Healthcare.



- Details about a patient (historical & current)-Query by MRN, Patient Name, Diagnosis, Date of Birth…
- Episodes of care (across visits)
- Identified patient information

- Patients matching criteria (symptoms, diagnosis, medication, prescriptions)
- De-identified patient information
- Medication effectiveness
- Early detection mechanisms

- Quality of Data captured the systems
- Cost of patients care
- Variations in Practice
- Quality Rating of the healthcare departments/business processes vis-à-vis benchmarks (HCAPPS scores are available across 5000 hospitals for free)

- Process turn around time, process optimization activities(when did the patient come in,where was the time spent in waiting, where are the bottle necks,…)
- Throughput Measurements(Discharge process turn around time,…)

[8]

Impact of Big Data on Healthcare

Predictive Analytics of Big Data in the domain of Healthcare can have a positive impact on a range of matters [2].

(1) **Cost Effectiveness:** Big data analytics can help reduce healthcare costs by different means, such as patient-outcome reimbursement, incorporation of Electronic Health Records, and getting rid of fraud and wastage in the Healthcare sector.

(2) **Convenient**: Patients can use evidence based care "on the go", and can even diagnose diseases online, through Online Medical Diagnosis Systems.

(3) **Preventive Care**: The highly individualized and real-time medical insights aid a wide range of patient care [3].

In the following topic, we shall see some of the Healthcare Analytics techniques used.

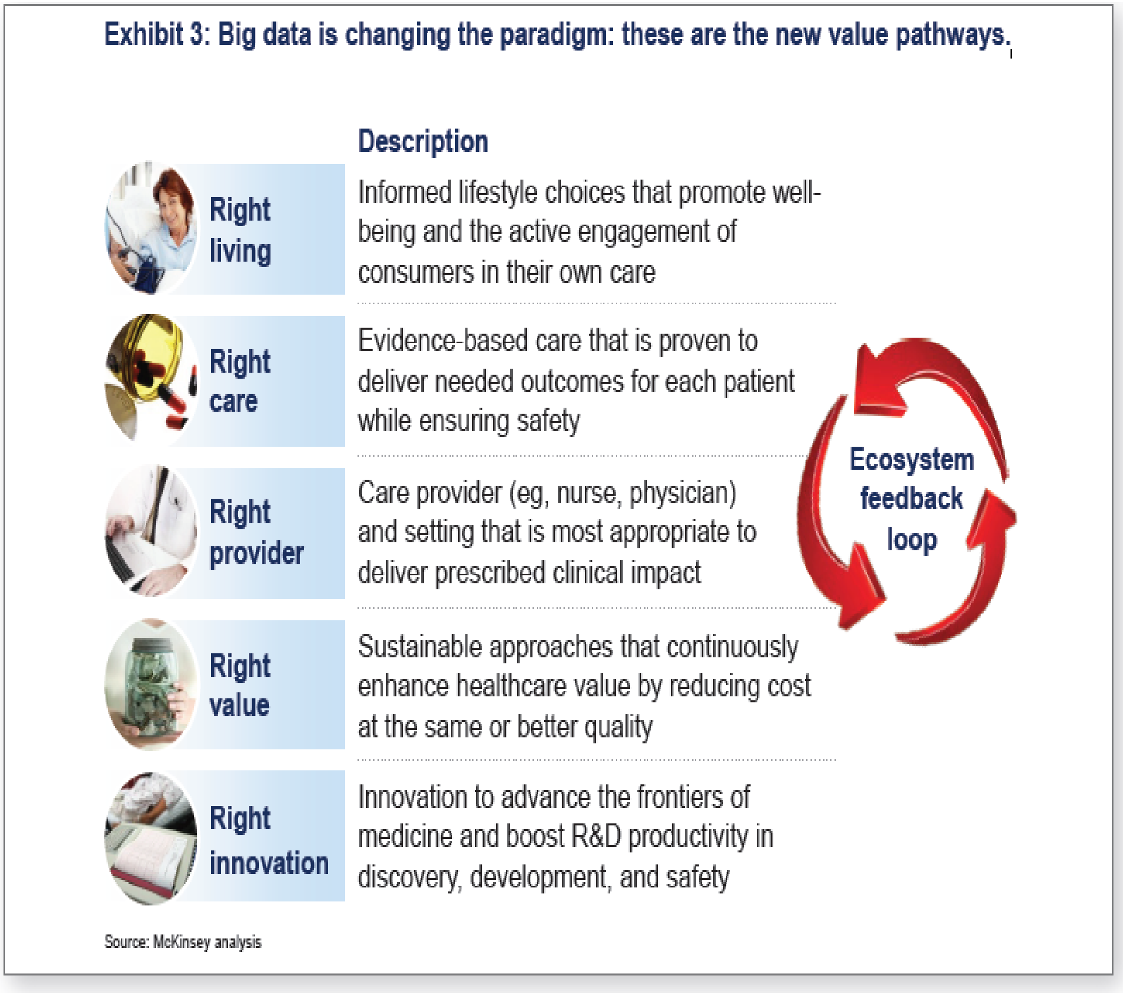


Exhibit 3: Big data is changing the paradigm: these are the new value pathways.

	Description
Right living	Informed lifestyle choices that promote well-being and the active engagement of consumers in their own care
Right care	Evidence-based care that is proven to deliver needed outcomes for each patient while ensuring safety
Right provider	Care provider (eg, nurse, physician) and setting that is most appropriate to deliver prescribed clinical impact
Right value	Sustainable approaches that continuously enhance healthcare value by reducing cost at the same or better quality
Right innovation	Innovation to advance the frontiers of medicine and boost R&D productivity in discovery, development, and safety

Ecosystem feedback loop

Source: McKinsey analysis

[2]

Common Analytics Techniques used in Healthcare:

Some common Big Data Analytics techniques used in the healthcare sector are discussed in this topic [6].

(1) **Support vector machines** are the supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. SVM constructs a hyper plane or set of hyper planes in high dimensional space which can be used for classification purposes.

(2) **Neural network** is a set of connected input/output unit, weight associated with each connection. The network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. The neural networks are used in many applications like pattern recognition problems, character recognition, and object recognition.

(3) **Naïve Bayes** is one of the successful classification methods. Classification is done based upon probability theory by computing prior Probability of the target attribute and conditional probability of remaining attributes. Naïve Bayes can't deal with continuous attributes so it is converted into discrete by using equal frequency discretization method.

(4) **Associative classification** is a new rule based approach that applies the methodology of association into classification. It adopts an exhaustive search algorithm like Apriori and FP growth to generate the class association rule. It selects the small set of high quality rules from large corpus of rules to construct an efficient classifier. It is a two-step process: first generate a large number of rules using any associative rule mining algorithm, and then several rule pruning techniques are used to generate an optimal rule set. There are two associative classification methods. Eager associative classification constructs a generalized model from a training data set before receiving any unknown instance for classification. Lazy approach does not previously build a generalized model from training data, but for each new instance to be classified, they process the stored training data samples.

Other techniques include rule induction and decision trees. Rule induction has a set of if-then-else rules, with two parts- a set of conditions (antecedents) and a set of associated results (consequents). It has the potential to use retrieved knowledge for prediction. Decision Tree is a

method of knowledge representation organized like a tree with branches and nodes. Every node is labelled with a class and branches coming out from an internal node have values of that node's attributes. This method is a common representation of instruction modelling. We also frequently use Genetic Algorithms (GA) and Nearest Neighbor Method (NNM). GAs are modelled on the process of genetic modification, alteration and natural selection, inspired by the observation of evolution in nature. The algorithm creates a number of solutions for the given problem. Next, weaker solutions are discarded and the rest are preserved. Solutions can overlap. A good solution is hybridized and the process is again repeated till we have the best possible solution. These are used with association rules and other internal formulations to formulate hypotheses about dependencies between variables here. NNMs are used mainly for classification. There is no pattern used to categorize data, data given as input is the pattern. NNM chooses the subset of input data which is the best possible fit and forecasts accordingly. NNM is used to check accuracy of heart disease diagnosis. Studies shows that it achieved 97.5% accuracy.

In the next section, I will discuss some applications of Big Data in Healthcare.

Applications of Big Data Analytics in Healthcare

Here, I illustrate some of the Applications of Big Data Analytics in Healthcare [6].

(1) General applications:

(a) Insurance fraud, Abuse detection, CRM decisions for in a healthcare organization, identifying effective treatment and best practices, providing patients with better and affordable services etc. There is a plethora of knowledge which can be gained from computerized health records but the vast amount of information stored makes it difficult. Application of data analytics on large medical data leads to discovery of new, useful and potentially lifesaving knowledge which would have otherwise been inert. Some current techniques are – telemedicine, Picture Archiving and Communication System (PACS), Digital Imaging and Communications in Medicine (DICOM), Electronic Medical Records (EMR) etc.

(b) Applications of Big Data Analytics in healthcare can be broadly classified into the following categories- Treatment Effectiveness, Healthcare management, Medical device industry, Pharmaceutical industry and System biology.

Now, I mention the Diagnostic Applications of Big Data Analytics in the Healthcare Sector.

(2) Existing Diagnostic Applications:

Here we look at various technologies and techniques that are used for diagnostic purposes in the healthcare sector:

(a) Current expert diagnostic systems that employ data analytics have a very low accuracy of prediction. This is because the techniques used are dependent on other records. They do not consider the patient's medical history and finally they are meant to be used only by domain experts and practitioners. To test this hypothesis, a data model was built to predict the blood sugar level of diabetic patient, which also considered the medical history. This was tested on a dataset from UCI Machine Learning repository, Washington University, St. Louis, Missouri. There were around 10,000 records per patient, each having information such as age, glucose level etc. There was 11.26% improvement in the accuracy of prediction and 4.37% reduction in number of false cases.

(b) A framework which uses density based multiple level clustering i.e., performing multiple iterations over the data collection has been proposed. This deals with the inherent sparseness and variable distribution of data. In each iteration, a different part of the data is analyzed and local clusters are identified for the set. The metric used is based on age, gender and examination history. By testing on a set of diabetic patients whose records are held by an Italian Health Centre, the framework progressively identified clusters of patients with advanced stages. The first iteration yielded the patients with routine tests and subsequent iterations identified those with specific tests. The cluster had good silhouette values and prediction was 90% accurate.

(c) A data driven Model Predictive Control which finds a suitable duration of Haemo-adsorption therapy for sepsis patients has been described. Here, therapy is applied in a non-continuous manner which saves 14% more patients than the usual method. According to this model, MPC is applied at each time point t to find the correct therapy for that time. If is recommended for that time, only one hour of treatment is administered. The patient is observed at the next point and so on. This was tested by looking through a population of non-survivor patients followed by a set of patients and temporal therapy to the patients. Finally, linear regression models for each variable are used on this training data. Non-continuous therapy cured around 41% with less than 12 hours of Haemo-adsorption therapy and in some cases, two hours was enough.

(d) The use of a hybrid Rough–Genetic algorithm model which implements the advantages of Rough Set as an efficient and powerful analysis tool to identify the most relevant attributes has been discussed. Firstly, Rough Sets can be used to discover important and relevant facts hidden in datasets and express them with decision rules of natural language, and these results (rules) from a Rough Set model are easily understood. Then, a Genetic Algorithm is used to optimize the rules induced using Rough Sets for classifying cases to test new medication for Hepatitis-C Virus (HCV) treatment. These algorithms encode a potential solution for a certain problem into a simple chromosome-like data structure, and then apply recombination operators to these data structures to preserve critical information. The experimental results obtained, show that the overall classification accuracy offered by the proposed Model is a dependable and superlative result.

(e) Predicting cardio vascular diseases based on Linear Discriminant Analysis of depression is possible. This also factors environmental variables like smoking, cholesterol levels, diabetes etc., for developing the prediction model. The dataset was obtained from the Korean National Health and Nutrition Examinations Survey (KNHANES) which had a total of 25,534 subjects, 335 depression patients were also included. Subjects were divided into training data and test data. Attributes considered are – sex, age, HDL cholesterol, total cholesterol, BP, smoking, diabetes and heart diseases. This information was applied to two equations for each patient. Equation 1 detects absence and equation 2 detects the presence of heart diseases. Classification is done depending on the higher value of the equations. Accuracy of FRS is 62.4% and for linear discriminant analysis, it is 69%.

(f) There is the Group Method of Data Handling (GMDH) for predictive modelling of healthcare data. GMDH is a family of inductive algorithms for computer based mathematical modelling of multi parametric datasets. To find the best solution, GMDH looks at various models estimated by the method of least squares. It increases the number of partial model components and finds a model with optimal complexity. This is measured by an external criterion, the minimal value of which indicates optimal complexity.

(g) Association Rule Summarizing techniques are used to detect the risk of Diabetes Mellitus. Common summarizing techniques are APRX – COLLECTION, RPGlobal, TopK and BUS. All techniques were applied to compress the original rule set of an electronic medical record to predict the risk of patients in the sub population. The most important differentiator between the techniques is the usage of a selection criterion to include a rule in the summary based on either expression of order or on the sub population covered by the rule. APRX and RPGlobal operate on expression while focussing on maximizing the compression, TopK and BUS operate on sub population with an aim to minimize redundancy. Association Rule Mining along with summarizing can help detect hidden clinical relationships and can also propose new patterns of conditions to redirect approaches for prevention, management and treatment. All four methods create reasonable summaries and each has its strengths.

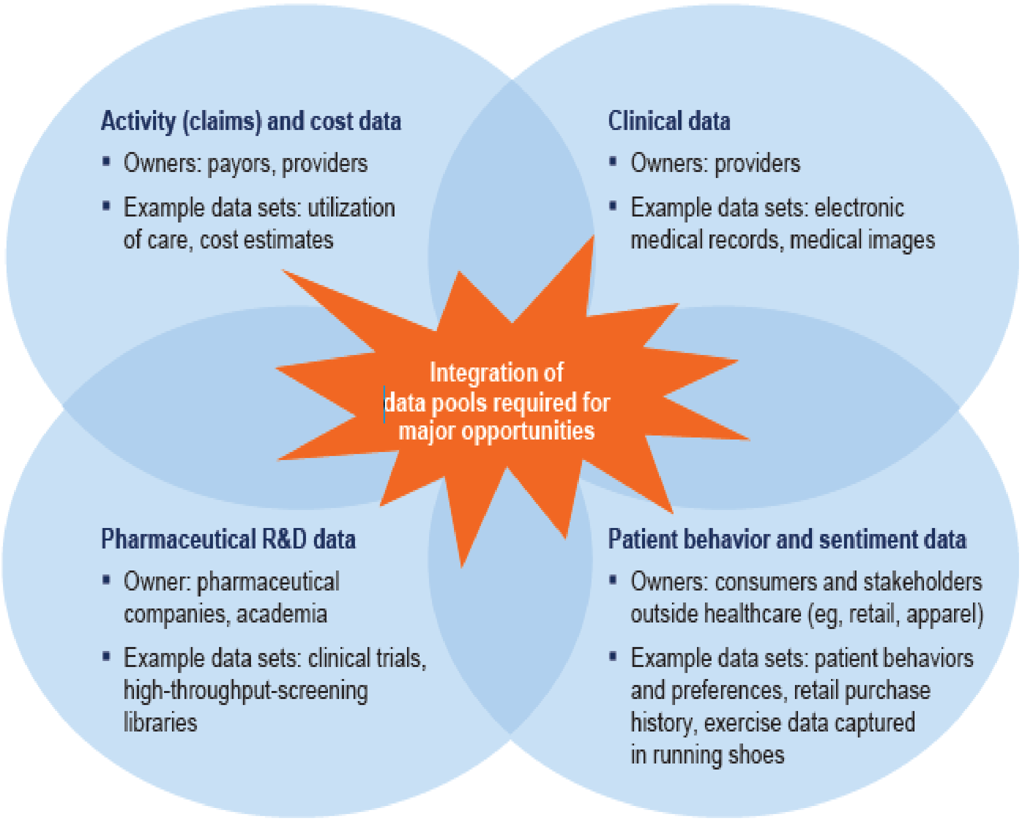
(h) Co-clustering can be used for diagnosis of heart diseases and also for detecting anomalies. Co-clustering acts as a powerful data analysis tool to diagnose heart disease and extract the data patterns of the datasets under test. Co-clustering, finds the subsets of rows in the dataset which are correlated with a subset of its columns. It is different from normal clustering that performs one way clustering such as k-means. On the other hand, in co-clustering simultaneous clustering of both row and columns happen. Co-clustering can produce a set of c column clusters of the original columns C and a set of r row clusters of original row instances R. Unlike other clustering algorithms, co-clustering also defines a clustering criterion and then optimizes it. In a nutshell, co-clustering finds out the subsets of rows and columns simultaneously of a data matrix using a specified criterion. From the summarization point of view, co-clustering provides significant benefits.

(i) Analytics techniques were used to show that Patient Characteristics are not associated with clinically important differential response to dapagliflozin. Baseline and early treatment response variable were selected and data analytics methods have been used to rank all variables which are linked with reduction in glycated hemoglobin (HbAic) at week 26. Generalized linear modelling was then implemented using an independent set of data values to figure out which variables were predictive of dapagliflozin-specific treatment response as compared with the response of treatment in the control aim of the study. Finally, the simplest model was chosen by meta-analysis of nine other trials. This approach helped in minimizing risk of type1 errors. From a very big set of data, 22 variables were used for generating the model as potentially predictive for dapagliflozin-specific reduction in HbA1c. Even though baseline HbA1c was the variable that is most strongly associated with reduction in HbA1c at the end of the study, baseline fasting plasma glucose (FPG) was found to be the only predictive dapagliflozin-specific variable in the model. Placebo adjusted treatment effect of dapagliflozin and metformin vs metformin only for a change in HbA1c from baseline which was found to be -0.65% at the average baseline FPG of 192.3 mg/dL. This output turned down by 0.32% for every SD [57.2 mg/dL] rise in baseline FPG. The baseline FPG effect was confirmed in the meta-analysis of 9 studies, but its quality was smaller. But no other variable was predictive of dapagliflozin-specific reduction in HbAic independently. This study successfully identified a baseline reproducible predictor of differential response to dapagliflozin. Even when, the predictor was shown as baseline FPG, its magnitude was small to suggest clinical usefulness in identifying patients who benefit from the treatment of dapagliflozin treatment uniquely. Till date, this is one of the limited examples of methodologies that identify the predictive variable within conventional clinical datasets, as generating during last-stage clinical trials.

(j) For pattern discovery in any type of Medical Data, we can use an enhanced k-means clustering algorithm. The approach is to group a given information set through a certain number of groups (expect k groups) that have been established beforehand. The principle idea is to characterize k centroids, one for each group. These centroids have to be set in a guile manner resulting in a distinctive area of diverse effects. In this way, the better decision is to place them as far as possible from one another. The subsequent step is to take each point within a given information set and co-partner it with the closest centroid until reaching a state where all the points have been associated with a group. Once the first stage is done, and an unanticipated aggregating is carried out automatically, we need to reconfigure k new centroids as barycenters of each group due to the last step. After producing these k new centroids, another binding must be established between the same centroids set and the closest new centroid. A cycle will be produced. As an after effect of this cycle, we may recognize that the k centroids will change their regulated areas and at the end of the day centroids will not change their positions anymore. K means needs improvement with the initial random selection of the centroids array. The first step is to calculate all of the existing elements that have the highest degree in the space; from there we can have an initial configuration of what the clusters should look like. On the second run, we eliminate all the centroids that are in a single cluster and select k clusters with the highest results of the similarity function to be taken as the real cluster centroids. This being done, we iterate on the rest of the data elements to see if the centroids are going to change This will be performed exactly like the original k-means with both the distance and similarity functions. This way, important and life-saving medical data patterns can be obtained.

(k) There is a cloud based healthcare application architecture titled “eHealth cloud” which uses a three tier architecture, each level having its own functionality. Tier-1 uses ria based client and enables the user to freely interact with the system. Secondly, the cloud server is simplified by using Amazon Simple DB. Finally, the logic layer between client and server contains the rules for the system. There are three types of users – patient, doctor and administrator each having their own interface. It employs various data analytics techniques for EMR.

Exhibit 2: Primary data pools are at the heart of the big-data revolution in healthcare.


Activity (claims) and cost data
▪ Owners: payors, providers
▪ Example data sets: utilization of care, cost estimates
Clinical data
▪ Owners: providers
▪ Example data sets: electronic medical records, medical images
Integration of data pools required for major opportunities
Pharmaceutical R&D data
▪ Owner: pharmaceutical companies, academia
▪ Example data sets: clinical trials, high-throughput-screening libraries
Patient behavior and sentiment data
▪ Owners: consumers and stakeholders outside healthcare (eg, retail, apparel)
▪ Example data sets: patient behaviors and preferences, retail purchase history, exercise data captured in running shoes

Source: McKinsey Global Institute analysis

Industry efforts to increase supply: Some firms and institutions with privileged access to big data

[2]

(3) Other Applications:

In addition to General and Diagnostic Applications, Big data Analytics has a variety of other applications in Healthcare.

(a) There exists a medical decision support system using the predictive modelling which predicts and prevents strain situations in hospitals and clinical facilities. There are 10 indicators of strain situations which are validated by professionals, one of which is Length of Stay (LOS). This makes use of several classification models and measures each one's performance using five metrics – Accuracy, Precision, Recall, Kappa Statistic and ROC. The dataset of 6,135 records between January and March 2012 was used. Exogenous values identified are – arrival time, age, tests etc. LOS is grouped into three types – < 289 minutes, 289 – 432 minutes, >432 minutes. Bayesian networks had the best precision (0.763), Kappa Statistic (0.36) and ROC (0.83). SVM had best accuracy (79.942%) and Recall (0.799).

(b) Similarly, there is a methodology using regression models to predict LOS of a new patient. Data is collected and formatted and the framework uses discrete event simulation to create new variables as per necessity. Identification of relevant variables is done using three approaches – hierarchical cluster analysis, attribute selection and principle component analysis. There are two linear models, both write the outcome as a sum of attribute values with appropriate weights assigned. The first model uses four values – Comp X-Ray, X-Ray, Echo and biology. The second model uses eight values – number of patients, AddressedBy , CAC, Echo, Scanner, X-Ray, AvisSpe, and Biology.

(c) Cross Entropy can be used for detection of anomalous behavior in Healthcare services. Due to the very large quantity of information and the increased cost in health services, faulty behavior may pass undetected and might be the reason of serious inefficiencies. Because the manual revision of data is costly and not convenient, other sources like data analysis and anomaly detection has to be used to generate effective quantity and anti-corruption controls. There is only one way in determining if information is anomalous. That is to compare it with the information that is provided by other agents. This method can be misleading at times. To compensate this problem, we have to discriminate information that is given by each agent by a group of risk factors. The risk group can be basically told as a set of individuals that share the same diagnosis and also the identical socio-economic characteristics. The analysis is done separately in each of the risk groups. For every agent and every risk group, the cross-entropy of the agent's information is calculated. So, each and every risk group will have a measure of how faulty the information is, provided by each agent. Also, as the risk groups size increases information is also available in big volume. The Columbian health system is one of few existing systems that use this mechanism. The results were found impressive and flexible with the fact that risk groups, variables can be defined with information from any year or any nation, and can be looked for anomalies.

Hence, these were the various Applications of Big Data Analytics in Healthcare. In the subsequent section, we shall see some of the challenges faced by Healthcare in the incorporation of Big Data Analytics.

Challenges

In this topic, I mention the various challenges pertaining to Big Data Analytics in Healthcare [6].

The challenges faced are:

(1) Data generated is enormous, and analyzing this huge corpus of "Big Data" is immensely challenging.

(2) Data obtained is usually Heterogeneous in nature and is from various sources which has an impact on data analysis and hence can't not be ignored.

(3) Also, interpretations of practitioners are in an unstructured language, making it difficult to mine such data.

(4) There are challenges w.r.t. knowledge integrity assessment – i.e. there is sometimes no guarantee that the obtained data is valid.

(5) There is also restricted access to raw inputs, and datasets have to be mostly manually generated.

(6) Improved data sharing between agencies has to be encouraged, along with the compression of data warehouses.

(7) Integration of very large Heterogeneous Medical Databases is hard, because medical data is diverse in nature and found across many databases in various formats. This requires integration of data stored in such a way that it is consistent.

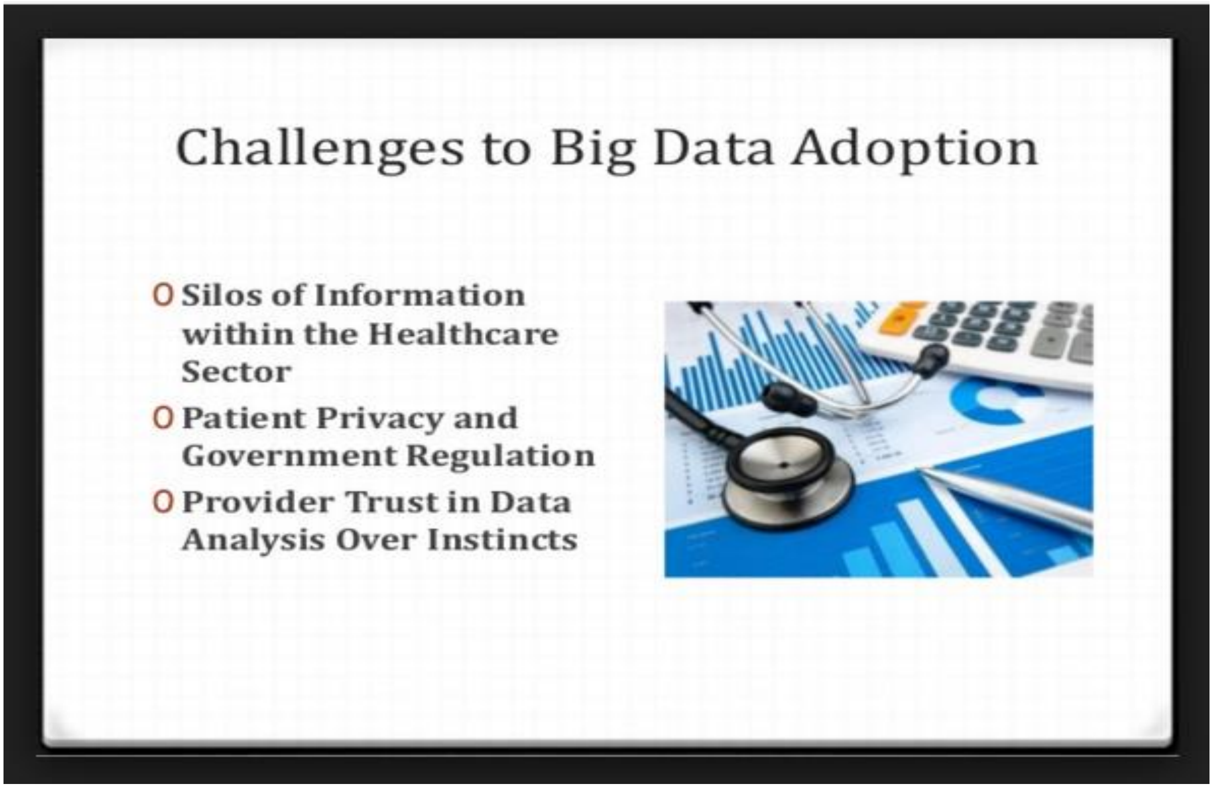
(8) Many a times, there is Redundancy due to overlapping of information in patient records and medical databases.

(9) Big data analytics in healthcare must be scalable and transparent.

(10) Privacy and security enablement is a major issue to be addressed [1].

(11) Data fragmentation is an obstacle to create a unified medical database.

(12) Ownership of Health information can be a loophole. [4]



Challenges to Big Data Adoption

- Silos of Information within the Healthcare Sector
- Patient Privacy and Government Regulation
- Provider Trust in Data Analysis Over Instincts

[7]

Conclusion

Big Data has only recently become a factor in the Healthcare Sector, and is being used in an array of ways, from analyzing Business practices in Healthcare to diagnosing Diseases and maintaining Health records. Big Data Analytics has the power to positively alter the way the Healthcare Industry incorporates technology, in order to gain insight into the large medical databases and clinical data, and thereby make informed medical decisions. A range of Analytics techniques and Algorithms are currently being proposed, which are steps in the right direction. Analytics can significantly help in discovering hidden patterns and relations in medical data, which could be potentially lifesaving. This massive amount of data is literally changing the way Medicine is practiced and the way the Healthcare Industry operates. Big Data Analytics in Healthcare still has a long, but encouraging journey ahead to positively impact Healthcare.

References

[1] Wullianallur Raghupathi and Viju Raghupati- "Big data analytics in healthcare: promise and potential"- Health Information Science and Systems 2014, 2:3

[2] Peter Groves, Basel Kayyali, David Knott, Steve Van Kuiken- "The 'big data' revolution in healthcare-Mckinsey 2013

[3] "How Big data impacts Healthcare"- Harvard Business review

[4] Michael Kassner- "Big data will enhance healthcare, but to whose benefit?"-TechRepublic

[5] Surya Nepal, Rajiv Ranjan, Kim-Kwang Raymond Choo- "Trustworthy Processing of Healthcare Big Data in Hybrid Clouds"- IEEE Cloud Computing, 2015-03

[6] Dhruv Madan Gopal, Aditya R, C Vishnu Kumar Reddy, Gautham S, Nagarathna N- "A Survey on Data Mining Applications, Techniques and Challenges in Healthcare"- Journal of Emerging Technologies and Innovative Research, 2015

[7] Jonathan Hannahs, David Hewitt, Chris Groves- "Big Data and the Healthcare Sector"- slideshare.net

[8] "Healthcare Big Data"- CMC Limited